

C-BiLDA

Extracting Cross-lingual Topics from Non-Parallel Texts by Distinguishing Shared from Unshared Content

Geert Heyman · Ivan Vulić · Marie-Francine Moens

Received: date / Accepted: date

Abstract We study the problem of extracting cross-lingual topics from non-parallel multilingual text datasets with partially overlapping thematic content (e.g., aligned Wikipedia articles in two different languages). To this end, we develop a new bilingual probabilistic topic model called comparable bilingual latent Dirichlet allocation (C-BiLDA), which is able to deal with such comparable data, and, unlike the standard bilingual LDA model (BiLDA), does not assume the availability of document pairs with identical topic distributions. We present a full overview of C-BiLDA, and show its utility in the task of cross-lingual knowledge transfer for multi-class document classification on two benchmarking datasets for three language pairs. The proposed model outperforms the baseline LDA model, as well as the standard BiLDA model and two standard low-rank approximation methods (CL-LSI and CL-KCCA) used in previous work on this task.

Keywords Cross-lingual text mining · multilingual topic modeling · multilinguality · comparable data · cross-lingual knowledge transfer · unsupervised modeling of text data · representation learning

1 Introduction

Cross-lingual text mining aims to induce and transfer knowledge across different languages to help applications such as cross-lingual information retrieval ([Levow et al, 2005](#); [Ganguly et al, 2012](#); [Vulić et al, 2013](#)), document classification ([Prettenhofer and Stein, 2010](#); [Ni et al, 2011](#); [Guo and Xiao, 2012a](#)), or cross-lingual annotation projection ([Zhao et al, 2009](#); [Das and Petrov, 2011](#); [van der Plas et al, 2011](#); [Kim et al, 2012](#); [Täckström et al, 2013](#); [Ganchev and Das, 2013](#)) in cases where translation and class-labeled resources are scarce or missing. In this article, we utilize probabilistic topic models to perform cross-lingual text mining. Probabilistic topic models are unsupervised generative models for representing document content in large document collections. Probabilistic topic models assume that every document is associated with a set of hidden variables, called topics, which determine how the words of the document were generated. Formally, a topic is a probability distribution

Geert Heyman · Ivan Vulić · Marie-Francine Moens
Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium
E-mail: geert.heyman@cs.kuleuven.be

over terms in a vocabulary. Informally, a topic represents an underlying semantic theme (Blei and McAuliffe, 2007). A representation of a document by such semantic themes has the advantage of being independent of both word-choice and language. Fitting a probabilistic topic model on a text collection is done by assigning the values to the hidden variables that best explain the data.

In monolingual settings the majority of text mining research using topic models is based on the probabilistic latent semantic analysis (pLSA) (Hofmann, 1999) or latent Dirichlet allocation (LDA) (Blei et al, 2003) models and its variants. Both are probabilistic models that take into account that word occurrences in the same document often belong to the same topic. This is done by associating a topic distribution to every document, rather than having a single topic distribution for the whole corpus. The models thus consist of two types of probability distributions: (1) distributions of topics over documents (further *per-document topic distributions*) and (2) distributions of words over topics (further *per-topic word distributions*). After learning the topic model on a training corpus, the obtained per-topic word distributions can be used to infer per-document topic distributions on unseen documents. The important difference between pLSA and LDA is that the latter takes the Bayesian approach for modelling the per-document topic distributions, i.e., the per-document topic distributions come from a Dirichlet-shaped prior distribution. pLSA in contrast uses point-estimates for the topic probabilities of documents, which makes it more vulnerable to overfitting. pLSA and LDA have found applications in document clustering, text categorization and ad-hoc information retrieval, but are not suited for cross-lingual text-mining since they were designed to work with monolingual data.

In multilingual settings, knowledge is mined from text by relying on machine-readable multilingual dictionaries or by using multilingual data. Since machine-readable dictionaries are not available for all languages pairs, the latter approach is more flexible. *Multilingual data* either refers to *parallel corpora* or *comparable corpora*. A parallel corpus is a collection of documents in different languages, where each document has a direct translation in the other languages. Hence, a parallel corpus is data-aligned at the sentence level. Parallel corpora are high-quality multilingual data resources, but they are not widely available for all language pairs and they are limited to a few narrow domains (e.g., the parliamentary proceedings of the Europarl corpus (Koehn, 2005)). Therefore, text mining from comparable corpora has gained interest over the last few years. A comparable corpus is a collection of documents with similar content which discusses similar themes in different languages, where documents in general are not exact translations of each other and are not strictly aligned at the sentence level. Unlike parallel corpora, comparable corpora by default comprise both shared and non-shared content.

A corpus built from Wikipedia using inter-wiki links to align content at the document level is a straightforward example of a comparable corpus, since the aligned article pairs may range from being almost completely parallel to containing non-parallel sentences. There are several other ways to acquire comparable corpora however. In the past years researchers have shown that comparable corpora can be automatically compiled from the Web. Utsuro et al (2002) construct comparable corpora with document alignments from English and Japanese news websites. To obtain a collection of similar documents they look at the dates of the articles and they rely on a machine translation tool to find document alignments. Talvensaari et al (2008) leverage the process of focussed crawling to obtain domain specific comparable corpora with paragraph alignments. The method was applied to gather comparable corpora in the genomics domain, and it was shown to be superior to a (general) parallel corpus in finding genomics related term translations. Apart from the resources we can find on the Web, organizations often possess domain specific corpora which allow to construct com-

parable corpora. In recent work for example, English and Chinese discharge summaries were used to create a comparable corpus in the healthcare sector (Xu et al, 2015). For even more approaches towards constructing document-aligned comparable data, we refer the interested reader to the relevant literature (Utiyama and Isahara, 2003; Tao and Zhai, 2005; Vu et al, 2009). While comparable corpora are typically cheaper, more abundant, more easily obtainable and more versatile than parallel corpora, they also constitute noisier and more challenging cross-lingual text mining environments.

Multilingual topic models such as bilingual LDA (BiLDA) (De Smet and Moens, 2009; Mimno et al, 2009) or Collaborative PLSA (C-PLSA) (Jiang et al, 2012) exploit the fact that the linked documents in multilingual corpora share content. These models assume that while the shared content is expressed with words from different vocabularies, the content can be represented in the same space of latent cross-lingual topics. Put differently, multilingual topic models learn cross-lingual topics which serve as a bridge between the different languages. The per-document word distributions constitute a language-independent document representation, while the language-specific information is modeled in per-topic word distributions. Topic models in this framework do not rely on sentence alignments, which makes them more robust to noisy data. However, the models assume that the topic distributions of linked documents are identical, which is not the case for comparable corpora.¹

Contributions. The main contribution of this article is a novel multilingual topic model specifically tailored to deal with non-parallel data. This model called *comparable bilingual LDA (C-BiLDA)* may be observed as an extension of the BiLDA model. However, unlike BiLDA, which assumes that two documents in an aligned document pair (e.g., a pair of aligned Wikipedia articles) share their topics completely (i.e., by modeling only one shared topic distribution), our new C-BiLDA model allows a document to elaborate more on certain topics than the document in the other language to which it is linked.

As another contribution, we show how to utilize our C-BiLDA model in the task of *cross-lingual knowledge transfer for multi-class document classification* for three language pairs. We show results on two datasets for a C-BiLDA-based transfer model which outscores LDA- and BiLDA-based transfer models previously reported in the literature (De Smet et al, 2011; Ni et al, 2011).

2 Related Work

One line of work in multilingual topic modeling explores multilingual topic models that are based on the premise of using readily available machine-readable multilingual dictionaries -if these are available at all- to establish links between content given in different languages which are in turn necessary to extract these latent cross-lingual topics (Boyd-Graber and Blei, 2009; Jagarlamudi and Daumé III, 2010; Zhang et al, 2010; Boyd-Graber and Resnik, 2010; Hu et al, 2014). In contrast, a more flexible multilingual topic modeling framework attempts to extract these latent topics solely on the basis of given multilingual data without any external resources at all. Due to its higher flexibility and scalability, our model is situated within this modeling framework. The standard multilingual model within this framework is called *bilingual LDA* (BiLDA) (De Smet and Moens, 2009; Ni et al, 2009; Platt et al, 2010;

¹ For instance, Wikipedia articles about Madrid in English and Spanish address many common topics such as “demographics”, “geography and location” or “climate”, while at the same time, only the Spanish article contains text (i.e., a non-shared topic) about “the emblems of the city”, and only the English article elaborates on “business schools” or “bohemian culture” in Madrid.

Zhang et al, 2013) or, by its extension to more than two languages, *polylingual LDA* (Mimno et al, 2009; Krstovski and Smith, 2013).²

All these models neglect one quite obvious fact - although dealing with comparable datasets which are inherently non-parallel and typically exhibit a degree of variance in their thematic/topical focuses, these models presuppose a perfect (or parallel) correspondence on extracted cross-lingual topics. More concretely, the models assume that the topic distributions of aligned documents are identical.

Aside from multilingual topic models, there are other approaches to mine cross-lingual word representations from multilingual corpora. Low rank methods and neural net models are two other commonly used approaches. Low rank methods use decompositions of co-occurrence matrices to find cross-lingual representations of words and/or documents. In multilingual text mining, cross-lingual latent semantic indexing (CL-LSI) and cross-lingual kernel canonical correlation analysis (CL-KCCA) are two established low rank methods. Given a parallel corpus, CL-LSI (Littman et al, 1998) concatenates the aligned document pairs and then applies LSI to find cross-lingual representations. CL-KCCA was proposed as an alternative to CL-LSI by Vinokourov et al (2002). After applying KCCA between the documents of source and target language respectively, semantic vectors for source and target language are constructed by projecting the respective document sets onto the k first correlation vectors. Each semantic vector corresponds to a cross-lingual topic. Documents can then be mapped to a cross-lingual representation by projecting their vector representation on the semantic vectors. Depending on its language, a document is projected on the semantic vectors of the source or target language. In the experiments of Vinokourov et al (2002), CL-KCCA with a linear kernel outperformed CL-LSI in both cross-lingual information retrieval and document classification.

The main focus of the neural net models lies on learning distributed word representations (dense real-valued vectors), which are shared across languages, by optimizing some criteria as a function of the data and the output of a neural network for which the words serve as input. Klementiev et al (2012) jointly train neural language models for two languages to induce shared cross-lingual distributed word representations. The neural language model learns distributed representations of words so that they can be used to predict the representation of the next word given the $n - 1$ previous words. To jointly learn the language models the multi-task learning setup of Cavallanti et al (2010) is used. Learning each vocabulary word in each language is considered a different task. To determine the degree of relatedness between two corresponding tasks, the approach requires the availability of *hard word alignments*, that is, links between words in parallel documents, where linked words are (part of) each others translations. Kočiský et al (2014) take a different approach and learn word representations that predict the representation of a word in the target language given $n - 1$ words in a parallel sentence in the source language. Both approaches build document representations simply as (weighted) averages of word representations. Instead of predicting a single word, Chandar et al (2014) learn to predict the bag-of-words representation of a target language sentence given the source language sentence.

Recently, Gouws et al (2014) have proposed a multilingual extension of the well-known word2vec models (Mikolov et al, 2013). Hermann and Blunsom (2014a,b) use a compositional vector model (CVM) to derive distributed representations for sentences and documents from distributed representations of words. The distributed representations are learned

² Without loss of generality, due to simplicity, we will restrict the presentation in the article to bilingual topic models.

by minimizing the energy between the distributed representation of parallel sentences.³ Soyer et al (2015) also use a composition function to compose words to phrases and sentences. They optimize both a bilingual objective and a monolingual objective. The bilingual objective is to minimize the energy between aligned sentence pairs. The monolingual objective aims to enforce that the energy between a sentence and a sub-phrase of the sentence is smaller than the energy between a sentence and a randomly sampled sub-phrase.

All these neural network based approaches actually need a strong bilingual signal given by (at least) a parallel corpus of a significant size (typically Europarl) in order to mine the knowledge from comparable datasets. In this work, we significantly alleviate the requirements, as we explicitly model both the shared and non-shared content in a document pair without the need for parallel data. In other words, unlike all previous work, our new model aims to extract *cross-lingual topics directly from non-parallel data by distinguishing between shared and unshared content, without any additional resources such as readily available bilingual lexicons or parallel data.*

3 Comparable Bilingual LDA

This section provides a full description of the newly designed C-BiLDA model. First, we define the standard BiLDA model, detect its limitations, and then introduce our new model which is able to handle comparable data. We present its core modeling premises, its relation to BiLDA, its generative story, and its training procedure by Gibbs sampling. In tab. 1 we summarize the notation used throughout the article.

3.1 Bilingual Topic Modeling

Assume that we possess an *aligned bilingual document corpus*, which is defined as $\mathcal{C} = \{d_1, d_2, \dots, d_D\} = \{(d_1^S, d_1^T), (d_2^S, d_2^T), \dots, (d_D^S, d_D^T)\}$, where $d_j = (d_j^S, d_j^T)$ denotes a pair of aligned documents in the source language L_S and the target language L_T , respectively. D is the number of aligned document pairs in the bilingual corpus. The goal of bilingual probabilistic topic modeling is to learn for the bilingual corpus a set of K latent cross-lingual topics $\mathcal{Z} = \{z_1, \dots, z_K\}$, each of which defines an associated set of words in both L_S and L_T (further denoted with superscripts S and T). A *bilingual probabilistic topic model* of a bilingual corpus \mathcal{C} is a set of multinomial distributions of words with values $P(w_i^S | z_k)$ and $P(w_i^T | z_k)$, $w_i^S \in V_S$, $w_i^T \in V_T$, where V_S and V_T are vocabularies associated with languages L_S and L_T . The aligned documents in a document pair need not be the exact translation of each other, that is, the corpus may be comparable and consist of documents which are only loosely equivalent to each other (e.g., Wikipedia articles in two different languages, news stories discussing the same event).

Each document, regardless of its language, may be uniformly represented as a mixture over induced latent cross-lingual topics using the probability scores $P(z_k | d_j)$ from per-document topic-distributions. This topic model-based representation allows for representing documents written in different languages in the same shared “topical” cross-lingual space. Topic modeling also enables learning the same cross-lingual representation for unseen data by utilizing the per-topic word distributions from the trained model to infer per-document topic distributions on the new data.

³ The energy between two vectors X and Y is defined as $\|X - Y\|^2$.

Documents, words and topics	
D	number of aligned document pairs
$d_j = (d_j^S, d_j^T)$	j -th pair of aligned documents
M_j and M_j^S	number of words in document pair d_j and source language document d_j^S respectively.
V_S	vocabulary of the source language
$ V ^S$	size of the vocabulary of the source language
w_l^S	l -th word of the source language vocabulary
w_{ji}^S	i -th word token of d_j^S
\mathbf{w}	vector with all word tokens in the corpus
l_{ji}	language corresponding to the i -th word token of document pair d_j
\mathbf{l}	vector with the same dimension as the word vector \mathbf{w} , where the i -th element is the language (L_S or L_T) of the i -th element in \mathbf{w}
\mathcal{Z}	set of latent cross-lingual latent topics
K	number of topics
z_k	k -th latent cross-lingual topic in \mathcal{Z}
z_{ji}	topic assigned to the i -th word token of d_j
z_{ji}^S	topic assigned to the i -th word token of d_j^S
\mathbf{z}	vector with all topic assignments in the corpus
Distribution parameters	
θ_j	topic distribution of the document pair d_j
θ_j^S	topic distribution of the source document d_j^S
δ_{jk}	probability that an occurrence of topic z_k in document pair d_j is assigned to a word in the source document
θ_{jk} and θ_{jk}^S	probability that a word token in document pair d_j and document d_j^S respectively is assigned to topic z_k
ϕ_k^S	distribution of the words in the source language for topic z_k
Hyperparameters	
α	parameter value of the symmetric Dirichlet prior on all θ_j
β	parameter value of the symmetric Dirichlet prior on all ϕ_k corresponding to topic z_k
χ_{jk}^S, χ_{jk}^T	parameter values for the Beta prior on all δ_{jk}
χ_{jk}	2-dimensional vector $\langle \chi_{jk}^S, \chi_{jk}^T \rangle$
Ω	set of all hyperparameters
Gibbs counters	
$n_{i,k}$	number of word tokens assigned to topic z_k in document pair d_j
$n_{j,k}^S$	number of word tokens assigned to topic z_k in document d_j^S
$n_{j,k,-i}$ or $n_{j,k,-i}^S$	number of word tokens assigned to topic z_k in document pair d_j or document d_j^S , excluding the word token at position i
$v_{k,l}^S$	number of times that word w_l^S is assigned to topic z_k
$v_{k,-i}^S$ or $v_{k,-i}^T$	number of times that word w_l^S is assigned to topic z_k , not counting the word token at position i in document d_j^S
$n_{j,i}^S$ or $n_{j,i,-i}^S$ or $v_{i,l}^S$ or $v_{i,-ji}^S$	replacing a subscript variable with a dot means summing over all possible values of that variable, e.g. $n_{j,i}^S = \sum_{k=1}^K n_{j,k}^S$

Table 1 A summary of the notation used throughout the article. For the language-specific notation we only show the notation for the source language (with the S superscript), while their counterpart in the target language is always obtained by replacing the S superscript with the T superscript.

The per-topic word and per-document topic distributions are learned in such a way so that they optimally explain the observed data. The exact calculation for this maximum likelihood criterion is intractable. Therefore, several approximate techniques have been proposed: Expectation-Maximization, variational Bayes, Gibbs sampling, etc. In this article we opt for the Gibbs sampling training technique, because of its popularity in literature and its ease of implementation. In its most general form, Gibbs sampling is a method to generate approximate samples from a joint distribution when directly sampling from the distribution is difficult or impossible. Starting from a random initial state, the Gibbs sampling algorithm generates a sample from the distribution of each variable in turn, conditioned on the values of all other variables in the current state (Bishop, 2006). Because the initialization of the sampling chain is done randomly, the samples in the beginning of the process are not representative. Therefore we start collecting samples when the chain reaches a stationary state (after the so-called *burn-in* period). Since successive samples are highly dependent, we only collect a sample for the variables every I -th value (e.g., every 20-th value).

3.2 Bilingual LDA

Bilingual LDA (Ni et al, 2009; De Smet and Moens, 2009; Mimno et al, 2009; Platt et al, 2010; Zhang et al, 2013) assumes that aligned documents have exactly the same per-document topic distributions. The graphical representation of BiLDA is given in fig. 1a. The model uses the same θ_j to model per-document topic distributions of documents in a pair. For each document pair d_j , a shared per-document topic distribution θ_j is sampled from a (symmetric) conjugate Dirichlet prior with K parameters $\alpha_1, \dots, \alpha_K$. Then, for each word position i in the source document of the current document pair d_j a cross-lingual topic z_k is sampled from θ_j (we denote this assignment by $z_{ji}^S = z_k$). Following that, a word is generated for every position i in document d_j^S by sampling from the multinomial distribution ϕ_k^S that corresponds to the topic z_k assigned to this position. Each word token w_{ji}^T from the target language side is also sampled following the same procedure, the only difference being that words are now sampled from the ϕ_k^T distributions. Note that words at the same positions in source and target documents in a document pair do not need to be sampled from the same latent cross-lingual topic. The only constraint imposed by the model is that the overall distributions of topics over documents in a document pair modeled by θ_j have to be the same. The validity of this assumption/constraint is dependent on the actual degree of thematic alignment between two coupled documents, and it perfectly fits only parallel documents which share all their topics. β is the parameter value of the symmetric Dirichlet prior on language-specific per-topic word distributions.

3.3 C-BiLDA: Extracting Shared and Non-Shared Topics

Modeling Assumptions. When one has to deal with a true comparable corpus, the assumption of “parallelism” exploited by BiLDA in its modeling premises is no longer valid, and it introduces several points of noise in the final output. As the same topics with the same proportions are used in both documents of a pair, there exists a clear discrepancy between learned topics and the actual content. In order to deal with the added difficulties caused by the “comparability” of the corpus and given document pairs, we extend the basic bilingual LDA model from sect. 3.2.

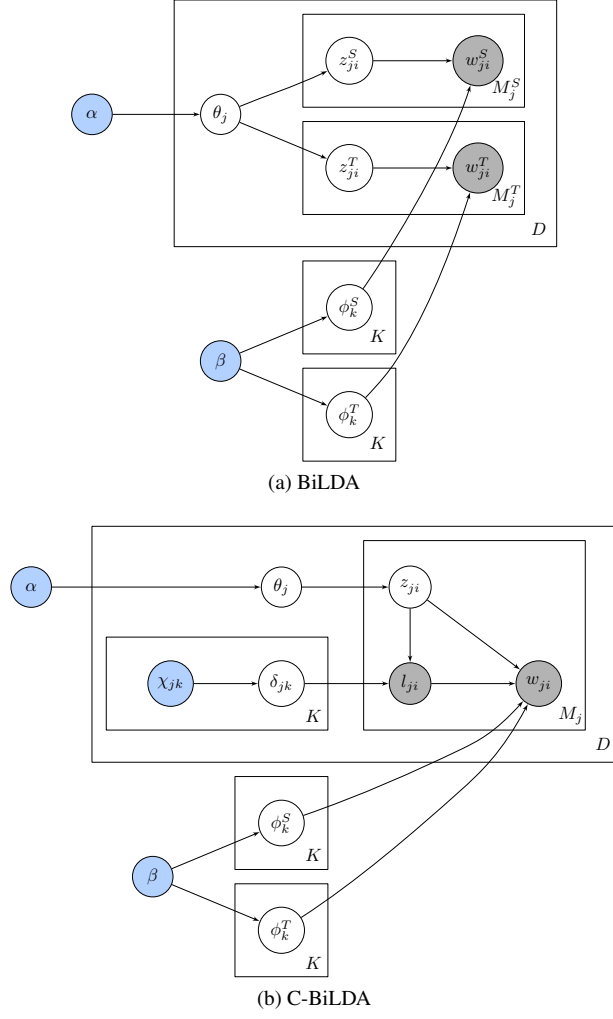


Fig. 1 Graphical representations of (a) BiLDA vs. (b) C-BiLDA in plate notation. BiLDA assumes that documents in an aligned document pair share all of their topics. Because of this assumption there is no need to represent the language l_{ji} of a topic occurrence. C-BiLDA on the other hand, allows the topic distributions of aligned documents to be different by assigning a language l_{ji} to every topic occurrence $z_{ji} = z_k$ depending on z_k : the source language is assigned to z_{ji} with probability δ_{jk} and the target language with probability $1 - \delta_{jk}$. M_j^S and M_j^T are the respective lengths of the source language document and the target language document in the j -th aligned document pair. M_j is the length of the document pair as a whole.

C-BiLDA allows a document to focus more on some topics than its counter part in the other language by modelling the probability that a topic occurrence in a document pair belongs to the source language. To this end we explicitly model the language l_{ji} for every word occurrence w_{ji} as an observed random variable and for each document introduce K parameters δ_{jk} describing the probability that a topic occurrence $z_{ji} = z_k$ in document pair d_j generates a word in the source language.

Algorithm 1: C-BiLDA: GENERATIVE STORY

```

initialize: (1) the total number of topics:  $K$ ;
              (2) the values for Dirichlet priors parameters  $\alpha$  and  $\beta$ ;
              (3) the values of all  $\chi_{jk}^S$  and  $\chi_{jk}^T$  (in fig. 1b we use  $\chi_{jk}$  as an abbreviation for  $\langle \chi_{jk}^S, \chi_{jk}^T \rangle$ )
sample  $K$  times  $\phi_k^S \sim \text{Dirichlet}(\beta)$ ;
sample  $K$  times  $\phi_k^T \sim \text{Dirichlet}(\beta)$ ;
for  $j \leftarrow 1$  to  $D$  do
  sample  $\theta_j \sim \text{Dirichlet}(\alpha)$ 
  sample  $K$  times  $\delta_{jk} \sim \text{Beta}(\chi_{jk}^S, \chi_{jk}^T)$ 
  for  $i \leftarrow 1$  to  $M_j$  do
    sample  $z_{ji} \sim \text{Multinomial}(\theta_j)$ 
    sample  $l_{ji} \sim \text{Bernoulli}(\delta_{jk})$ , with  $z_{ji} = z_k$ 
    if  $l_{ji} = 1$  then
      sample  $w_{ji} \sim \text{Multinomial}(\phi_k^S)$ , with  $z_{ji} = z_k$ 
    else
      sample  $w_{ji} \sim \text{Multinomial}(\phi_k^T)$ , with  $z_{ji} = z_k$ 

```

Generating the Data. Fig. 1b shows the plate representation of C-BiLDA. As in the BiLDA generative process, all topics of a document pair are drawn from the same distribution θ_j , but source and target documents can have a preference to certain topics. After generating a topic $z_{ji} = z_k$ from θ_j , we sample the language l_{ji} associated with this topic occurrence from a Bernoulli distribution with δ_{jk} as the probability of success. We place a Beta prior with parameter values χ_{jk}^S and χ_{jk}^T on all δ_{jk} . These values can be interpreted as psuedo-counts for observing topic z_k in the source/target document of document pair d_j respectively. After sampling a topic-language pair, a word is generated in the same way as in the BiLDA model, that is, by sampling from the word distribution of the sampled topic in the sampled language. The distributions θ_j , ϕ_k^S , ϕ_k^T and corresponding hyperparameters α and β are the same as in BiLDA (see sect. 3.2). Alg. 1 summarizes the generative story of C-BiLDA.

Relation with BiLDA. In its original formulation BiLDA looks quite different from C-BiLDA. This is because with the BiLDA assumptions, it is not necessary to model the language of a word as a random variable. However, we can represent BiLDA exactly like C-BiLDA with the exception of using a single δ_j (representing the probability that any topic will generate a word in the source document) per document, instead of using K δ_{jk} variables per document (one for each topic)⁴. Therefore, C-BiLDA allows a document to focus more on a particular topic than its counterpart or, in the extreme case, to contain topics that do not occur in its counterpart. The added flexibility also has a downside since it increases the risk of overfitting the data. By setting an appropriate prior on all δ_{jk} variables, we can avoid that C-BiLDA learns models that are too complex. By setting the prior values of $\chi_{j1}^S, \dots, \chi_{jK}^S$ to the same value and similarly for the values of $\chi_{j1}^T, \dots, \chi_{jK}^T$, we make the a priori assumption that the topic distributions for source and target document are identical (like in BiLDA). In our experiments we set $\chi_{jk}^S = \frac{1}{2} \chi_m M_j^S$ and $\chi_{jk}^T = \frac{1}{2} \chi_m M_j^T$. The document sizes M_j^S and M_j^T are observed, so only the value of χ_m must be set manually. The higher the value of χ_m , the more weight we give to the prior assumption that the source and target document topic distributions are the same, and the closer the C-BiLDA relates to BiLDA.

⁴ By writing out the joint probability conditioned on all language assignments l_{ji} , one can check that these formulations are indeed equivalent.

Algorithm 2: GIBBS SAMPLING FOR C-BiLDA: AN OVERVIEW

```

Algorithm gibbsSampler()
  repeat
    | sampleTopics();
  until burn-in criterion satisfied
  repeat
    for  $i \leftarrow 1$  to  $I$  do
      | sampleTopics();
    end
    collect a sample: estimate  $\theta_{jk}, \delta_{jk}, \phi_{kl}^S, \phi_{kl}^T$  from the current topic assignments using eq.
    (6)-(9);
  until enough samples collected
  estimate the posteriors of  $\theta_{jk}, \delta_{jk}, \phi_{kl}^S, \phi_{kl}^T$  by averaging over the collected samples;
Procedure sampleTopics()
  foreach word token in the corpus do
    update/estimate the probability to assign the word token to one of the cross-lingual topics
    conditioned on all other variables (for C-BiLDA apply eq. (1));
    sample a new topic assignment for the word token;
  end

```

Training. To infer the values of the unobserved variables, we utilize Gibbs sampling (Geman and Geman, 1984; Bishop, 2006). Note that from the vector of all topic assignments: \mathbf{z} together with the observed word and language variables, all other latent variables can be derived. The values of all $\theta_j, \delta_{jk}, \phi_k^S$ and ϕ_k^T can be integrated out of the formulas and calculated afterwards. All other variables are observed (the word tokens in the bilingual corpus and their corresponding languages) or are hyperparameters that have to be set in advance (α , all χ_{jk} and β). Therefore one iteration of the Gibbs sampling procedure estimates the topic assignments for each word in turn by sampling their probability distribution conditioned on all other variables. The high-level Gibbs sampling procedure for C-BiLDA is shown in alg. 2, below we derive the necessary update formulas for z_{ji} .

$$\begin{aligned}
& P(z_{ji} = z_k | w_{ji} = w_l, l_{ji}, \mathbf{z}_{-ji}, \mathbf{w}_{-ji}, \mathbf{l}_{-ji}, \Omega) \\
&= \frac{P(z_{ji} = z_k, w_{ji} = w_l, l_{ji}, \mathbf{z}_{-ji}, \mathbf{w}_{-ji}, \mathbf{l}_{-ji}, \Omega)}{P(w_{ji} = w_l, l_{ji}, \mathbf{z}_{-ji}, \mathbf{w}_{-ji}, \mathbf{l}_{-ji}, \Omega)} \\
&= \frac{P(w_{ji} = w_l, \mathbf{w}_{-ji} | z_{ji} = z_k, l_{ji}, \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \Omega) \cdot P(z_{ji} = z_k, l_{ji}, \mathbf{z}_{-ji}, \mathbf{l}_{-ji} | \Omega)}{P(\mathbf{w}_{-ji} | l_{ji}, \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \Omega) \cdot P(w_{ji} | l_{ji}, \Omega) \cdot P(l_{ji}, \mathbf{z}_{-ji}, \mathbf{l}_{-ji} | \Omega)} \\
&\propto P(w_{ji} = w_l | z_{ji} = z_k, l_{ji}, \mathbf{z}_{-ji}, \mathbf{w}_{-ji}, \mathbf{l}_{-ji}, \Omega) \cdot P(z_{ji} = z_k | l_{ji}, \mathbf{z}_{-ji}, \mathbf{l}_{-ji} | \Omega) \\
&\propto P(w_{ji} = w_l | z_{ji} = z_k, l_{ji}, \mathbf{z}_{-ji}, \mathbf{w}_{-ji}, \mathbf{l}_{-ji}, \Omega) \cdot \frac{P(z_{ji} = z_k, l_{ji}, \mathbf{z}_{-ji}, \mathbf{l}_{-ji} | \Omega)}{P(l_{ji}, \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \Omega)} \\
&\propto P(w_{ji} = w_l | z_{ji} = z_k, l_{ji}, \mathbf{z}_{-ji}, \mathbf{w}_{-ji}, \mathbf{l}_{-ji}, \Omega) \cdot \frac{P(l_{ji}, \mathbf{l}_{-ji} | z_{ji} = z_k, \mathbf{z}_{-ji}, \Omega) \cdot P(z_{ji} = z_k, \mathbf{z}_{-ji} | \Omega)}{P(\mathbf{l}_{-ji} | \mathbf{z}_{-ji}, \Omega) \cdot P(l_{ji} | \Omega) \cdot P(\mathbf{z}_{-ji} | \Omega)} \\
&\propto P(w_{ji} = w_l | z_{ji} = z_k, l_{ji}, \mathbf{z}_{-ji}, \mathbf{w}_{-ji}, \mathbf{l}_{-ji}, \Omega) \cdot P(l_{ji} | z_{ji} = z_k, \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \Omega) \cdot P(z_{ji} = z_k | \mathbf{z}_{-ji}, \Omega) \\
&\propto \begin{cases} E[\theta_{jk} | \mathbf{z}_{-ji}, \alpha] \cdot E[\delta_{jk} | \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \chi_{jk}^S, \chi_{jk}^T] \cdot E[\phi_{kl}^S | \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \mathbf{w}_{-ji}, \beta] & \text{if } l_{ji} = S \\ E[\theta_{jk} | \mathbf{z}_{-ji}, \alpha] \cdot (1 - E[\delta_{jk} | \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \chi_{jk}^S, \chi_{jk}^T]) \cdot E[\phi_{kl}^T | \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \mathbf{w}_{-ji}, \beta] & \text{if } l_{ji} = T \end{cases} \quad (1)
\end{aligned}$$

$$\text{with } E[\theta_{jk} | \mathbf{z}_{-ji}, \alpha] = \frac{n_{j,k,-i} + \alpha}{n_{j,-i} + K\alpha} \quad (2)$$

$$\text{and } E[\delta_{jk} | \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \chi_{jk}^S, \chi_{jk}^T] = \frac{n_{j,k,-i}^S + \chi_{jk}^S}{n_{j,k,-i}^S + \chi_{jk}^S + \chi_{jk}^T} \quad (3)$$

$$\text{and } E[\phi_{kl}^S | \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \mathbf{w}_{-ji}, \beta] = \frac{v_{k,-ji}^S + \beta}{v_{k,-ji}^S + |V|^S \cdot \beta} \quad (4)$$

$$\text{and } E[\phi_{kl}^T | \mathbf{z}_{-ji}, \mathbf{l}_{-ji}, \mathbf{w}_{-ji}, \beta] = \frac{v_{k,-ji}^T + \beta}{v_{k,-ji}^T + |V|^T \cdot \beta} \quad (5)$$

The final estimates of the posteriors of θ_{jk} , δ_{jk} , ϕ_{kl}^S and ϕ_{kl}^T are calculated by estimating their posteriors for every sample that is taken using equations (6)-(9) and then taking the average of these estimates over all samples.

$$E[\theta_{jk} | \mathbf{z}, \alpha] = \frac{n_{j,k} + \alpha}{n_{j,\cdot} + K\alpha} \quad (6)$$

$$E[\delta_{jk} | \mathbf{z}, \mathbf{l}, \chi_{jk}^S, \chi_{jk}^T] = \frac{n_{j,k}^S + \chi_{jk}^S}{n_{j,k}^S + \chi_{jk}^S + \chi_{jk}^T} \quad (7)$$

$$E[\phi_{kl}^S | \mathbf{z}, \mathbf{l}, \mathbf{w}, \beta] = \frac{v_{k,l}^S + \beta}{v_{k,\cdot}^S + |V|^S \cdot \beta} \quad (8)$$

$$E[\phi_{kl}^T | \mathbf{z}, \mathbf{l}, \mathbf{w}, \beta] = \frac{v_{k,l}^T + \beta}{v_{k,\cdot}^T + |V|^T \cdot \beta} \quad (9)$$

Inferring topic distributions. For certain tasks (e.g., information retrieval) it is necessary to infer a topic model on unseen data. Inferring the model actually denotes calculating per-document topic distributions on unseen documents based on the output of the trained model. Again, we use Gibbs sampling to approximate the distribution, but now we use the per-topic word distributions learned after training. Therefore, we only update the n counters. Furthermore, the inference is done monolingually, that is one language at a time. The updating formula for the source language L_S is:

$$P(z_{ji}^S = k | w_{ji}^S = w_l, \mathbf{z}_{-ji}^S, \mathbf{w}_{-ji}^S, \alpha, \beta) \propto E[\theta_{jk} | \mathbf{z}_{-ji}^S] \cdot E[\phi_{kl}^S | \text{training data}]$$

$$\text{with } E[\theta_{jk} | \mathbf{z}_{-ji}^S] = \frac{n_{j,k,-i}^S + \alpha}{n_{j,-i}^S + K\alpha} \quad (10)$$

Where the n counters count topic assignments for *unseen documents* and $E[\phi_{kl}^S | \text{training data}]$ is the estimate of ϕ_{kl}^S on the training data.

4 Knowledge Transfer via Cross-Lingual Topics for Document Classification

The per-topic word distributions of multilingual topic models can be used for a variety of tasks. One application is to map the distributions to per-word distributions, e.g., $p(z_k | w_i)$ or $p(w_i, w_j)$. This results in a type of distributed word representation for w_i , which in turn can be used to find word associations and/or extract translation pairs, etc. (Vulić et al, 2011). In this article, we demonstrate the utility of our new C-BiLDA model on yet another task: cross-lingual document classification, as it is a well-established cross-lingual task that gives insight into cross-lingual text mining models and their ability to learn semantically-aware

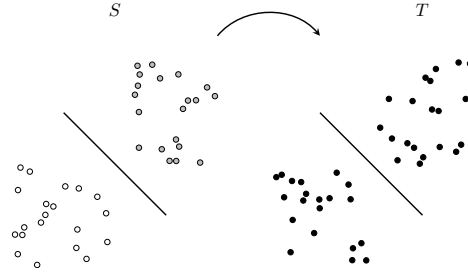


Fig. 2 An intuition behind cross-lingual knowledge transfer for document classification. White and gray circles denote labeled examples, while black circles denote unlabeled examples.

document representations.

Problem Definition. Cross-lingual document classification (CLDC) is the task of assigning class labels to documents written in the target language given the knowledge of the labels in the source language (Bel et al, 2003; Gliozzo and Strapparava, 2006). It starts from a set of labeled documents in the (resource-rich) source language, and unlabeled documents in the (resource-poor) target language. The objective is to learn a classification model from the labeled documents of the source language and then *transfer this knowledge* to the target language and apply it in the classification model for the target language documents (see fig. 2 for a more intuitive presentation).

Previous Work. Early approaches to the problem of CLDC tried to utilize automatic machine translation tools to translate all the data from S to T , which effectively reduced the problem to monolingual classification (Bel et al, 2003; Fortuna and Shawe-Taylor, 2005; Olsson et al, 2005; Rigutini et al, 2005; Ling et al, 2008; Wei and Pal, 2010; Duh et al, 2011; Wan et al, 2011). Other approaches rely on machine translation tools along with multi-view learning (Amini et al, 2009; Guo and Xiao, 2012a) or co-training techniques (Wan, 2009; Amini and Goutte, 2010; Lu et al, 2011). However, machine translation tools may not be freely available for many language pairs, which limits the portability of these models. In addition, translating all the text data is often time-consuming and expensive.

Another line of prior work aims to induce cross-lingual representations for documents given in different languages, which enables the knowledge transfer for CLDC using the shared language-independent feature spaces. A plethora of CLDC models have been proposed (Gliozzo and Strapparava, 2006; Prettenhofer and Stein, 2010; Pan et al, 2011; Wang et al, 2011; Klementiev et al, 2012; Guo and Xiao, 2012b; Xiao and Guo, 2013b,a; Hermann and Blunsom, 2014b; Chandar et al, 2014), but all these models again assume that parallel corpora or external translation resources are readily available to induce these cross-lingual shared representations.

Finally, in order to overcome these issues, another line of recent work (De Smet et al, 2011; Ni et al, 2011) operates in a minimalist setting; it aims to learn these *shared cross-lingual representations directly from non-parallel data* without any other external resources such as high-quality parallel data or machine-readable bilingual lexicons. These approaches train a multilingual topic model (e.g., BiLDA) on *comparable data* to induce topical representations of documents, and use per-document topic distributions as classification features. In this article, we show that for this setup the application of C-BiLDA instead of BiLDA

Table 2 Statistics of the Wikipedia and Europarl training sets.

	Wikipedia-dataset			Europarl-dataset		
	EN-ES	EN-FR	EN-IT	EN-ES	EN-FR	EN-IT
$ V^S $	29,201	27,033	23,346	33,444	33,574	33,552
$ V^T $	27,745	20,860	31,388	36,839	34,538	36,092
#Doc-pairs	18,672	18,911	18,898	9,415	9,428	9,461

leads to a better performance.

Knowledge Transfer via Latent Topics. The idea is to take advantage of the cross-lingual representations by means of latent cross-lingual topics. First a topic model (e.g., BiLDA or C-BiLDA) is trained on a bilingual training corpus (e.g., Wikipedia). Following that, given a CLDC task, with a labeled set of documents in the source language and an unlabeled document collections in the target language, one uses the trained topic model to infer the cross-lingual representations by means of per-document topic distributions for each (previously unseen) document. Each document is then taken as a data instance in the classification model and the features are defined as probabilities coming from per-document topic distributions. The value of each feature of an instance (e.g., a document d_j^S) is the probability of the corresponding topic z_k in the document: $P(z_k | d_j^S)$ (see sect. 3.1). Finally, one is free to choose any classifier (e.g., Maximum Entropy, Naive Bayes, Support Vector Machine) to perform classification.

5 Experimental Setup

Training Datasets. To train the topic models on a comparable corpus, we use the training dataset of De Smet et al (2011) for the same CLDC task (while the dataset from (Ni et al, 2011) is not publicly available). It consists of three bilingual corpora with aligned Wikipedia articles in three language pairs: English-Spanish (EN-ES), English-French (EN-FR), and English-Italian (EN-IT). The datasets were collected from Wikipedia *dumps*, and the alignment between articles in a pair was obtained by following the inter-lingual Wikipedia links. Stop words were removed, and only words that occur at least 5 times were retained. To show the influence of the degree of parallelism in the training data, we also train C-BiLDA and BiLDA on a parallel corpus extracted from Europarl. The resulting dataset uses the same language pairs as the Wikipedia dataset and the processing was done in the same way. Tab. 2 lists statistics of the training datasets.

CLDC Datasets. We test our models by performing CLDC on two different datasets. We run the trained topic models on these test datasets, that is, we infer the per-document topic distributions, which are then used for training and testing a classifier. In all experiments, we regard English as the resource-rich language and learn class labels for test documents in the other 3 target languages (ES/FR/IT) with labels removed from their documents.

The first dataset again comes from De Smet et al (2011). It was constructed using Wikipedia. The dataset for each language pair contains up to 1,000 Wikipedia articles (which are not present in the training sets) annotated with 5 high-level labels/classes: *book* (books), *film* (films), *prog* (computer programming), *sport* (sports) and *video* (video games). Since not every Wikipedia in every language contains the same number of articles, sometimes less than 1,000 articles for each class was crawled from Wikipedia *dumps*. For more details about the dataset construction, we refer the interested reader to (De Smet et al, 2011).

Table 3 Number of documents in the CLDC datasets.

	Wikipedia-dataset					RCV1/RCV2-dataset			
	<i>book</i>	<i>film</i>	<i>prog</i>	<i>sport</i>	<i>video</i>	<i>MCAT</i>	<i>CCAT</i>	<i>GCAT</i>	<i>ECAT</i>
EN	1,000	1,000	1,000	1,000	1,000	7,441	12,934	7,216	2,409
ES	1,000	1,000	263	1,000	1,000	9,694	30	1,997	1279
FR	1,000	1,000	592	1,000	1,000	5,878	65	20,987	3,070
IT	1,000	1,000	290	1,000	764	7,553	263	1,520	3,664

To compare the BiLDA and C-BiLDA models on a larger corpus we constructed a second dataset from the Reuters corpora RCV1/RCV2 (Lewis et al, 2004). The dataset contains up to 30,000 documents per language. Since our training dataset does not include the English-German language pair that was used by Klementiev et al (2012), we could not reuse their dataset. We constructed the dataset with the procedure from Klementiev et al. for the three language pairs in our training dataset: we use the top-level category labels that are assigned to the documents: *CCAT* (Corporate/Industrial), *ECAT* (Economics), *GCAT* (Government/Social), *MCAT* (Markets); and only consider documents with a single top-level topic. Similar to Klementiev et al (2012), we sample randomly from the original RCV1/RCV2 corpora, but for the language pairs in our training dataset. The documents from both datasets were preprocessed in the same manner as in the training datasets. Tab. 3 displays the size of the CLDC datasets.

Models in Comparison. We test the ability of our new C-BiLDA model to transfer the knowledge needed for cross-lingual document classification, and compare it to other topic modeling approaches for knowledge transfer previously reported in the literature. The models in comparison are:

1. *CL-LSI-TR*. A CLDC model based on CL-LSI (Littman et al, 1998). In order to come up with uniform cross-lingual representations, it combines each aligned pair of documents into an artificial “merged document”, keeping no language-specific information. On the merged documents (monolingual) LSI is applied. The rank reduced term-document matrix (where the new rank is equal to the number of topics) is then used to project the documents in the cross-lingual space in which we train the classifier.
2. *CL-KCCA-TR*. This model is based on the CL-KCCA model of Vinokourov et al (2002). The semantic vectors of the source/target language are used to project documents of the source/target respectively in the cross-lingual space in which we train the classifier. Like Vinokourov et al (2002) we use a linear kernel.
3. *LDA-TR*. This was the baseline model in (De Smet et al, 2011). Similar to CLLSI-TR it combines each aligned pair of documents into an artificial “merged document”. The merged documents are then used to train a monolingual LDA (Blei et al, 2003) model, which is then inferred on the test documents. Per-document topic distributions are then used as features for classification.
4. *BiLDA-TR*. This is the best scoring model in De Smet et al (2011); Ni et al (2011), which also significantly outperformed models relying on machine translation tools and bilingual lexicons (Ni et al, 2011). BiLDA is trained on aligned documents, and then inferred on test data. Per-document topic distributions are again used as features for classification (see sect. 4).
5. *C-BiLDA-TR- χ_m* , with $\chi_m \in \{0.125, 0.25, 0.5, 1, 2\}$. As for BiLDA, we train C-BiLDA on aligned document pairs to obtain per-document topic distributions. We use different

values of χ_m (recall from sect. 3.3 that χ_m determines the values of the prior parameters χ_{jk}^S and χ_{jk}^T).

Parameters. Following prior work, we use a Support Vector Machine (SVM) for classification with all transfer models. For SVM, we employ the SVM-Light package⁵ (Joachims, 1999) with default parameter settings. Investigating other choices for classifiers, as well as different classifier settings is beyond the scope of this article. All models are trained for different number of topics K , ranging from 20 to 300 in steps of 20. CL-LSI was implemented using the *truncated svd* module of scikit-learn⁶ (Pedregosa et al, 2011). For CL-KCCA we used KCCA package by Hardoon et al (2004). The regularization parameter κ was set using the method proposed in Hardoon et al (2004).

Hyperparameters α and β in LDA and BiLDA are set to the standard values according to (Steyvers and Griffiths, 2007): $\alpha = 50/K$ and $\beta = 0.01$. In case of C-BiLDA, we show the results for different values of χ_m : $\{0.125, 0.25, 0.50, 1, 2\}$. The higher the χ_m value, the higher the influence of the priors on δ_{jk} . The topic models have been trained by Gibbs sampling. As the burn-in criterion, we check if the relative difference of the perplexity between two iterations is smaller than a predefined small threshold value (we use 0.0001 in all training procedures). After the burn-in period, we gather samples every $I = 20$ iterations. The total number of iterations (including the burn-in period) is set to 1000. Perplexity is a measure for the likelihood of the data for a given statistical model. The perplexity on a corpus \mathcal{C} for a statistical model \mathcal{M} is defined as:

$$\text{perplexity}(\mathcal{C}|\mathcal{M}) = \exp\left(-\frac{\sum_{j=1}^D \prod_{i=1}^{M_j} \log(p(w_{ji}|\mathcal{M}))}{\sum_{j=1}^D M_j}\right)$$

Evaluation Metrics. For each category, precision is calculated as the number of correctly labeled documents divided by the total number documents that have been labeled this way. Recall is defined as the number of correctly labeled documents divided by the actual number of documents with that label given by the ground truth. Precision and recall are then combined into balanced *F-1 scores*. We calculate macro F-1 scores by taking the average of the F-1 scores over all categories and all K s. For BiLDA and C-BiLDA, we also report the perplexities on the training datasets. Perplexity measures how well a statistical model fits the data.

6 Results and Discussion

Perplexity and Comparability. In this paragraph we analyse the perplexity of C-BiLDA and BiLDA on the different training datasets. Tab. 4 shows average perplexity scores of C-BiLDA and BiLDA models trained on the parallel Europarl corpora and the comparable Wikipedia corpora. The perplexity scores confirm our hypothesis that BiLDA is better fit for modeling parallel data, while C-BiLDA is tailored for more divergent, comparable data.

In tab. 4 we also show the difference in perplexity between the two models: perplexity BiLDA – perplexity C-BiLDA. We expect this difference to be an indicator of the degree of comparability of a multilingual corpus. The larger the difference between the perplexity of BiLDA and the perplexity of C-BiLDA models, the less parallelism we expect to find in the data because we expect C-BiLDA to model non-parallelism in a better way. The results in tab.

⁵ <http://svmlight.joachims.org/>

⁶ <http://scikit-learn.org/>

Table 4 Perplexity scores of the BiLDA and C-BiLDA models and their difference (the perplexity score of BiLDA minus the perplexity score of C-BiLDA) on the Wikipedia training datasets averaged across the number of topics and χ_m values. From the perplexity scores and the difference in perplexity scores of C-BiLDA and BiLDA we can rank the training datasets according to their level of parallelism.

	Wikipedia			Europarl		
	<u>EN-ES</u>	<u>EN-FR</u>	<u>EN-IT</u>	<u>EN-ES</u>	<u>EN-FR</u>	<u>EN-IT</u>
perpl. BiLDA	2827	2544	3042	1564	1391	1600
perpl. C-BiLDA	2787	2504	2839	1581	1402	1615
perpl. BiLDA - perpl. C-BiLDA	40	40	203	-17	-11	-15

4 confirm this hypothesis, since on the comparable Wikipedia dataset the difference in perplexity values is higher than for the parallel Europarl datasets. The results also indicate that the EN-IT Wikipedia dataset is less parallel than the EN-FR and EN-ES Wikipedia datasets, since the difference in perplexity is larger. For the Wikipedia datasets the overall perplexity is higher for EN-ES than for EN-FR. This is an indication that the latter is the Wikipedia dataset with the most parallelism.

CLDC Task. Tab. 5 summarizes the performance in the CLDC task of the transfer models (TRs) with representations trained on Wikipedia. F-1 scores are macro-averaged over different category labels and averaged over different K s. Tab. 5 also ranks the training datasets in their degree of comparability, based on the perplexity analysis in the previous paragraph. Fig. 3 shows how F-1 scores fluctuate on the Reuters test dataset across different K values for BiLDA and C-BiLDA with $\chi_m = 2$. From these results we may observe several interesting phenomena:

(i) The difference between LDA on one side and BiLDA and C-BiLDA is very profound. While all these transfer models are based on the same principle, and use per-document topic distributions to provide language-independent document representations, separating the vocabularies and training a true bilingual topic model on individual documents from aligned pairs (instead of removing all language information from the corpus) is clearly more beneficial for the CLDC task. Similar findings have been reported for cross-lingual information retrieval (Jagaramudi and Daumé III, 2010; Vulić et al, 2013) and word translation identification (Vulić et al, 2011, 2015).

(ii) Also the difference between the low-rank approximation methods (CL-LSI, CL-KCCA) on one side and C-BiLDA and BiLDA is profound. An explanation for this may be that the use of priors in the probabilistic framework is a robust way to deal with the non-parallelism in comparable corpora.

(iii) When comparing BiLDA with the C-BiLDA transfer models we see that the C-BiLDA models generally perform better. For the CLDC task on the Wikipedia test set, both the C-BiLDA transfer models and the BiLDA transfer model have good F-1 scores, indicating that the models learn representations that are well suited for the Wikipedia test set. The differences between the C-BiLDA and BiLDA models are not so profound as for the Reuters test set. After performing a qualitative inspection of the topic distributions, we conclude there is a clean mapping between the topics we learned from our training data and the categories of the Wikipedia dataset. The representations of the categories of the Reuters dataset on the other hand, are more spread out across topics. In the latter case it is more important to have more clean/coherent topics overall. Therefore, we conclude that C-BiLDA is able to learn “cleaner” per-topic word distributions.

(iv) We observe that for the language pair with the least comparable training data, the C-BiLDA transfer models perform better than the BiLDA model and that the C-BiLDA models

Table 5 CLDC with representations trained on Wikipedia. Average F-1 scores on the Wikipedia and Reuters test sets with 8 different transfer models for each language pair. Average F-1 is calculated by macro-averaging the F-1 scores over all category labels and all Ks. Classifier is SVM. + sign indicates a better F-1 score of a C-BiLDA-TR when compared to the baseline models. The best F-1 scores per language pair are shown in bold.

	EN-ES		EN-FR (most parallel)		EN-IT (least parallel)	
TR-Model	Wiki	Reuters	Wiki	Reuters	Wiki	Reuters
CL-LSI	31.17	27.59	28.44	35.35	26.79	21.06
CL-KCCA	14.03	14.12	24.03	24.28	10.21	8.91
LDA	32.84	7.55	34.65	10.08	30.99	26.86
BiLDA	81.46	39.74	76.88	45.30	78.36	45.22
C-BiLDA ₂	81.51+	40.77+	76.61	45.63+	78.83+	46.21+
C-BiLDA ₁	80.64	40.27+	74.47	44.92	79.27+	45.66+
C-BiLDA _{0.5}	80.83	39.70	76.19	45.30	79.09+	46.76+
C-BiLDA _{0.25}	79.71	40.41+	75.03	44.48	79.06+	46.08+
C-BiLDA _{0.125}	79.91	40.41+	75.37	44.02	78.85+	45.60+

with lower χ_m values perform best (recall that a lower χ_m value in fact implies assigning less weight to the a priori *parallel document pair assumption*, see sect. 3.3). On the other hand, for the EN-FR language pair we observe that the difference between C-BiLDA and BiLDA is less profound and that the higher values for the χ_m parameter perform best. This intuition underpinned by the reported results reveals a link between the comparability of the training data and the performance of the BiLDA model and the C-BiLDA models with different χ_m . (v) From fig. 3 we conclude that the difference between the C-BiLDA transfer model with $\chi_m = 2$ and the BiLDA transfer model are consistent for the lower topic values. For the higher topic values performance begins to drop. This illustrates previously mentioned overfitting problems. More topics lead to more model parameters, for C-BiLDA even more so than BiLDA.

Further Discussion. One may argue that capturing additional phenomena in the data (e.g., document pairs with non-parallel document distributions) leads to an added complexity in the model design. However, the increased design complexity is justified by the need to capture the properties of non-parallel data. Consequently, the final scores in the CLDC task further justify the requirement for a more complex topic model which is better aligned to the given data.

We have reported that the priors placed on the δ_{jk} variables have significant influence on the quality of the learned topics. Their values should be high enough to avoid overfitting, though low enough to take into account non-parallelism (i.e., non-shared content) in document pairs. It may be too time/resource consuming to explore what values for the χ priors are appropriate by trying different values and finding out which work best. One approach we intend to investigate in future work is to treat the hyperparameters as random variables that are learned from the data just like the other parameters. McCallum et al (2009) have successfully applied this approach to the α hyperparameters for monolingual LDA.

So far we have not talked about the minimum the degree of comparability between the corpora in order to learn any useful bilingual knowledge. This is a difficult question in general. For C-BiLDA in particular, the document pairs may exhibit low comparability in case the following conditions hold for the document collection as a whole: (1) the document collection should contain enough cross-lingual information, this means that as the comparability between document pairs goes down, the size of the document collection should go up

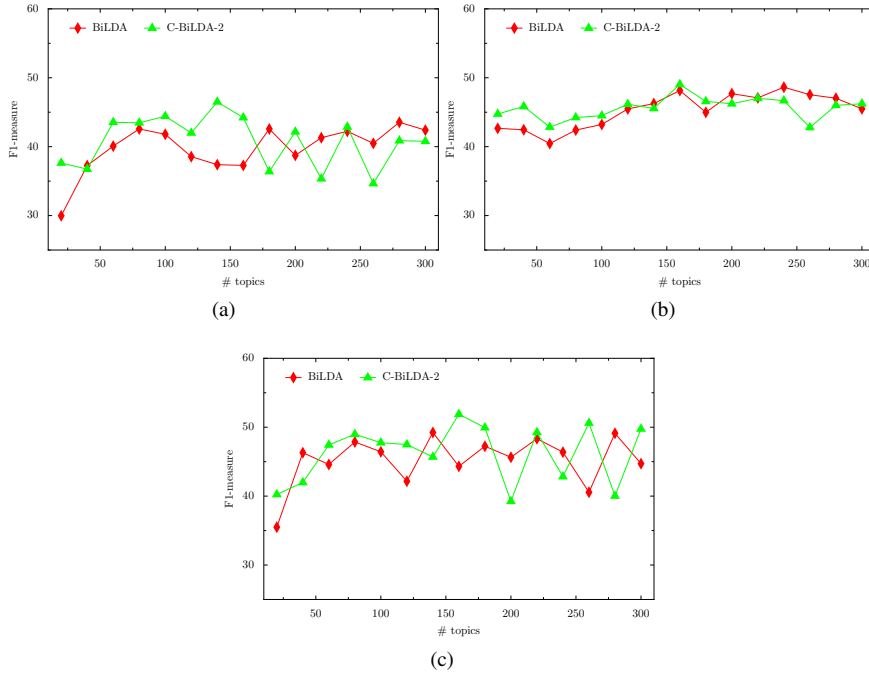


Fig. 3 The average F1-scores for a varying amount of topics for the BiLDA transfer model and the C-BiLDA transfer model with $\chi_m = 2$ on the CLDC task with the Reuters dataset: EN-ES (a), EN-FR (b) and EN-IT (c).

accordingly; (2) if a theme often reoccurs in the documents of the source language, it should often occur in the documents of the target language. This requirement can be fulfilled by ensuring that the document collection is restricted to a limited domain.

Besides the CLDC task, we believe that the proposed C-BiLDA model and the idea of distinguishing between shared and unique content in related documents may find further application in other tasks. One interesting application is tackled in (Paul and Girju, 2009), where they analyze cultural differences between speakers of the same language across different countries and cultures. A similar idea applied to the analysis of ideological differences is discussed in (Ahmed and Xing, 2010). Another interesting future application is the analysis of differences between Twitter and traditional media (Zhao et al, 2011). The C-BiLDA model and its extensions in future research may be utilized to induce different views on the same subjects/concepts/topics given in different languages and/or in different media, as well as to extract language-specific concepts from blogs, forums, tweets and online discussions.

7 Conclusions

We have studied the problem of extracting cross-lingual topics from non-parallel data. In this article, we have presented a new bilingual probabilistic topic model called comparable bilingual LDA (C-BiLDA) which is able to distinguish between shared and unshared content in aligned document pairs to learn more coherent cross-lingual topics. We have

demonstrated the utility of C-BiLDA in performing the knowledge transfer for cross-lingual document classification for three language pairs, where our model has outperformed the standard bilingual LDA model (BiLDA) on two benchmarking datasets, indicating that distinguishing between shared and unique content in document pairs leads to better per-topic word distributions when training on non-parallel data. Like other topic models, C-BiLDA can be used in a variety of other natural language processing and information retrieval tasks.

C-BiLDA is completely data-driven and does not require a machine-readable bilingual dictionary or high-quality parallel data. Furthermore it does not make any language specific assumptions. C-BiLDA's wide applicability in terms of input data makes it an excellent model for learning representations in under-resourced languages and language pairs, as well as in domains with specific terminology for which high-quality (multilingual) data is often not available.

Compliance with Ethical Standards

Funding: The research presented in this article has been carried out in context of the **SCATE** (SBO-130047) research project financed by the (Flemish) agency for Innovation through Science and Technology (IWT).

Conflict of Interest: The authors declare that they have no conflict of interest.

References

- Ahmed A, Xing EP (2010) Staying informed: Supervised and semi-supervised multi-view topical analysis of ideological perspective. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1140–1150
- Amini MR, Goutte C (2010) A co-classification approach to learning from multilingual corpora. *Machine Learning* 79(1-2):105–121
- Amini MR, Usunier N, Goutte C (2009) Learning from multiple partially observed views - an application to multilingual text categorization. In: Proceedings of the 23rd Annual Conference on Advances in Neural Information Processing Systems (NIPS), pp 28–36
- Bel N, Koster CHA, Villegas M (2003) Cross-lingual text categorization. In: Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), pp 126–139
- Bishop CM (2006) *Pattern Recognition and Machine Learning* (Information Science and Statistics). Springer-Verlag New York, Inc.
- Blei DM, McAuliffe JD (2007) Supervised topic models. In: Proceedings of the 21st Annual Conference on Advances in Neural Information Processing Systems (NIPS), pp 121–128
- Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022
- Boyd-Graber J, Blei DM (2009) Multilingual topic models for unaligned text. In: Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI), pp 75–82
- Boyd-Graber J, Resnik P (2010) Holistic sentiment analysis across languages: Multilingual supervised latent Dirichlet allocation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 45–55
- Cavallanti G, Cesa-Bianchi N, Gentile C (2010) Linear algorithms for online multitask classification. *The Journal of Machine Learning Research* 11:2901–2934

- Chandar S, Lauly S, Larochelle H, Khapra MM, Ravindran B, Raykar VC, Saha A (2014) An autoencoder approach to learning bilingual word representations. In: Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems (NIPS)
- Das D, Petrov S (2011) Unsupervised part-of-speech tagging with bilingual graph-based projections. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), pp 600–609
- De Smet W, Moens MF (2009) Cross-language linking of news stories on the Web using interlingual topic modeling. In: Proceedings of the CIKM 2009 Workshop on Social Web Search and Mining (SWSM@CIKM), pp 57–64
- De Smet W, Tang J, Moens MF (2011) Knowledge transfer across multilingual corpora via latent topics. In: Proceedings of the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp 549–560
- Duh K, Fujino A, Nagata M (2011) Is machine translation ripe for cross-lingual sentiment classification? In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), pp 429–433
- Fortuna B, Shawe-Taylor J (2005) The use of machine translation tools for cross-lingual text mining. In: Proceedings of the ICML 2005 KCCA Workshop (KCCA)
- Ganchev K, Das D (2013) Cross-lingual discriminative learning of sequence models with posterior regularization. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1996–2006
- Ganguly D, Leveling J, Jones G (2012) Cross-lingual topical relevance models. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING), pp 927–942
- Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6(6):721–741
- Gliozzo AM, Strapparava C (2006) Exploiting comparable corpora and bilingual dictionaries for cross-language text categorization. In: Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and the 21st International Conference on Computational Linguistics (ACL-COLING)
- Gouws S, Bengio Y, Corrado G (2014) Bilbowa: Fast bilingual distributed representations without word alignments. *Deep Learning Workshop, Conference on Neural Information Processing Systems (NIPS)*
- Guo Y, Xiao M (2012a) Cross language text classification via subspace co-regularized multi-view learning. In: Proceedings of the 29th International Conference on Machine Learning (ICML)
- Guo Y, Xiao M (2012b) Transductive representation learning for cross-lingual text classification. In: Proceedings of the 12th IEEE International Conference on Data Mining (ICDM), pp 888–893
- Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: An overview with application to learning methods. *Neural computation* 16(12):2639–2664
- Hermann KM, Blunsom P (2014a) Multilingual distributed representations without word alignment. In: Proceedings of the International Conference on Learning Representations (ICLR)
- Hermann KM, Blunsom P (2014b) Multilingual models for compositional distributed semantics. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), pp 58–68

- Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI), pp 289–296
- Hu Y, Zhai K, Eidelman V, Boyd-Graber JL (2014) Polylingual tree-based topic models for translation domain adaptation. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), pp 1166–1176
- Jagarlamudi J, Daumé III H (2010) Extracting multilingual topics from unaligned comparable corpora. In: Proceedings of the 32nd Annual European Conference on Advances in Information Retrieval (ECIR), pp 444–456
- Jiang Y, Liu J, Li Z, Lu H (2012) Collaborative PLSA for multi-view clustering. In: Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE, pp 2997–3000
- Joachims T (1999) Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A (eds) *Advances in Kernel Methods - Support Vector Learning*, MIT Press, chap 11, pp 169–184
- Kim S, Toutanova K, Yu H (2012) Multilingual named entity recognition using parallel data and metadata from Wikipedia. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL), pp 694–702
- Klementiev A, Titov I, Bhattarai B (2012) Inducing crosslingual distributed representations of words. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING), pp 1459–1474
- Koehn P (2005) Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the 10th Machine Translation Summit (MT SUMMIT), pp 79–86
- Kočiský T, Hermann KM, Blunsom P (2014) Learning bilingual word representations by marginalizing alignments. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL), pp 224–229
- Krstovski K, Smith DA (2013) Online polylingual topic models for fast document translation detection. In: Proceedings of the Workshop on Statistical MT
- Levow GA, Oard DW, Resnik P (2005) Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management* 41(3):523–547
- Lewis DD, Yang Y, Rose TG, Li F (2004) RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research* 5:361–397
- Ling X, Xue GR, Dai W, Jiang Y, Yang Q, Yu Y (2008) Can Chinese Web pages be classified with English data source? In: Proceedings of the 17th International Conference on World Wide Web (WWW), pp 969–978
- Littman M, Dumais ST, Landauer TK (1998) Automatic cross-language information retrieval using Latent Semantic Indexing. In: Chapter 5 of *Cross-Language Information Retrieval*, Kluwer Academic Publishers, pp 51–62
- Lu B, Tan C, Cardie C, K Tsou B (2011) Joint bilingual sentiment classification with unlabeled parallel corpora. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), pp 320–330
- McCallum A, Mimno DM, Wallach HM (2009) Rethinking lda: Why priors matter
- Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space
- Mimno D, Wallach H, Naradowsky J, Smith DA, McCallum A (2009) Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 880–889
- Ni X, Sun JT, Hu J, Chen Z (2009) Mining multilingual topics from Wikipedia. In: Proceedings of the 18th International World Wide Web Conference (WWW), pp 1155–1156
- Ni X, Sun JT, Hu J, Chen Z (2011) Cross lingual text classification by mining multilingual topics from Wikipedia. In: Proceedings of the 4th International Conference on Web

- Search and Web Data Mining (WSDM), pp 375–384
- Olsson JS, Oard DW, Hajič J (2005) Cross-language text classification. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp 645–646
- Pan J, Xue GR, Yu Y, Wang Y (2011) Cross-lingual sentiment classification via bi-view non-negative matrix tri-factorization. In: Proceedings of the 15th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD), pp 289–300
- Paul MJ, Girju R (2009) Cross-cultural analysis of blogs and forums with mixed-collection topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1408–1417
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830
- van der Plas L, Merlo P, Henderson J (2011) Scaling up automatic cross-lingual semantic role annotation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), pp 299–304
- Platt JC, Toutanova K, Yih WT (2010) Translingual document representations from discriminative projections. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 251–261
- Prettenhofer P, Stein B (2010) Cross-language text classification using structural correspondence learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp 1118–1127
- Rigutini L, Maggini M, Liu B (2005) An EM based training algorithm for cross-language text categorization. In: Proceedings of the 2005 ACM International Conference on Web Intelligence (WIC), pp 529–535
- Soyer H, Stenetorp P, Aizawa A (2015) Leveraging monolingual data for crosslingual compositional word representations
- Steyvers M, Griffiths T (2007) Probabilistic topic models. *Handbook of Latent Semantic Analysis* 427(7):424–440
- Täckström O, McDonald R, Nivre J (2013) Target language adaptation of discriminative transfer parsers. In: Proceedings of the 14th Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp 1061–1071
- Talvensaari T, Pirkola A, Järvelin K, Juhola M, Laurikkala J (2008) Focused web crawling in the acquisition of comparable corpora. *Information Retrieval* 11(5):427–445
- Tao T, Zhai C (2005) Mining comparable bilingual text corpora for cross-language information integration. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp 691–696
- Utiyama M, Isahara H (2003) Reliable measures for aligning Japanese-English news articles and sentences. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), pp 72–79
- Utsuro T, Horiuchi T, Chiba Y, Hamamoto T (2002) Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on WWW news sites. Springer
- Vinokourov A, Cristianini N, Shawe-Taylor JS (2002) Inferring a semantic representation of text via cross-language correlation analysis. In: *Advances in neural information processing systems*, pp 1473–1480

- Vu T, Aw AT, Zhang M (2009) Feature-based method for document alignment in comparable news corpora. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp 843–851
- Vulić I, Moens MF (2015) Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In: In Proceedings of the 38th Annual ACM SIGIR Conference on Research and Development in Information Retrieval
- Vulić I, De Smet W, Moens MF (2011) Identifying word translations from comparable corpora using latent topic models. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), pp 479–484
- Vulić I, De Smet W, Moens MF (2013) Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora. *Information Retrieval* 16(3):331–368
- Vulić I, Smet WD, Tang J, Moens M (2015) Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing and Management* 51(1):111–147
- Wan C, Pan R, Li J (2011) Bi-weighting domain adaptation for cross-language text classification. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), pp 1535–1540
- Wan X (2009) Co-training for cross-lingual sentiment classification. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL), pp 235–243
- Wang H, Huang H, Nie F, Ding C (2011) Cross-language Web page classification via dual knowledge transfer using nonnegative matrix tri-factorization. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), pp 933–942
- Wei B, Pal CJ (2010) Cross lingual adaptation: An experiment on sentiment classifications. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp 258–262
- Xiao M, Guo Y (2013a) A novel two-step method for cross language representation learning. In: Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems (NIPS), pp 1259–1267
- Xiao M, Guo Y (2013b) Semi-supervised representation learning for cross-lingual text classification. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp 1465–1475
- Xu Y, Chen L, Wei J, Ananiadou S, Fan Y, Qian Y, Chang EIC, Tsujii J (2015) Bilingual term alignment from comparable corpora in english discharge summary and chinese discharge summary. *BMC Bioinformatics*
- Zhang D, Mei Q, Zhai C (2010) Cross-lingual latent topic extraction. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL), pp 1128–1137
- Zhang T, Liu K, Zhao J (2013) Cross lingual entity linking with bilingual topic model. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), pp 2218–2224
- Zhao H, Song Y, Kit C, Zhou G (2009) Cross language dependency parsing using a bilingual lexicon. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL), pp 55–63
- Zhao WX, Jiang J, Weng J, He J, Lim EP, Yan H, Li X (2011) Comparing twitter and traditional media using topic models. In: Proceedings of the 33rd European Conference

on Advances in Information Retrieval (ECIR), pp 338–349